Arcada Working Papers 7/2017
ISSN 2342-3064
ISBN 978-952-5260-85-4

City Event Identification from Instagram Data using Word Embedding and Topic Model Visualization*

Shuhua Liu[i] and Patrick Jansson[ii]

Abstract

In this study, we explore unsupervised methods and models for identifying city events from Instagram data. We combine the use of event keywords with expansions based on word embeddings to identify semantically related terms and seed events, then retrieve event-relevant posts. We apply LDA topic modelling method to the set of relevant posts to a given class of event to discover clusters of targeted events. The system helps us to shed light on the presence or absence of certain type of city events on Instagram posts in the region.

Keywords: word embeddings, LDA topic modelling, event extraction, city events, social media

1 INTRODUCTION

The world's urban population reached more than 4.11 billion in 2017, which amounts to 54.7% of our total population[1]. Many city authorities and city planners face various challenges in planning future developments, in deploying, maintaining and optimizing urban infrastructure. Understanding city systems and urban dynamics has never been so important and crucial for smooth functioning of modern cities and regions.

Social media is gaining more and more attention from city planners, city authorities and urban researchers as potentially useful source of information for developing better understanding of city dynamics. In this study, we explore state of the art NLP methods and techniques to detect city events from Instagram data.

There exists a large amount of literature on event detection and extraction from social media, especially Twitter. Atefeh and Khreich (2015) presented a survey of techniques for event detection in twitter. Sakaki et al. (2010) proposed a probabilistic model to de-

* Funding from Helsinki Region Urban Research Program (http://www.helsinki.fi/kaupunkitutkimus/) and Arcada Foundation TUF (http://tuf.arcada.fi) are gratefully acknowledged.
[i] Arcada University of Applied Sciences, Dept. of Business Management and Analytics, [Shuhua.liu@arcada.fi]
[ii] Arcada University of Applied Sciences, Dept. of Business Management and Analytics, [Patrick.jansson@arcada.fi]
[1] Source: http://www.worldometers.info/world-population/

tect the location of earthquake and typhoon occurrences. Ritter et al. (2012) pioneered an open domain event extraction and categorization system for Twitter. LDA topic modelling was applied to detect topic clusters combined with manual inspection of the clusters types. Li et al. (2014) worked on accurate extraction of major life events and formulation of fine-grained description of users' current life events based on their published tweets.

Ritter et al. (2015) proposed a seed-based weakly supervised approach for the detection and extraction of computer security events. Jain et al. (2016) studied the detection of civil unrest events from Twitter, adopting a weakly supervised, multi-faceted event recognition approach. Their system automatically learns event indicators and identifies Tweets about an event using a single seed event phrase capturing defining characteristics of the event, without having any prior knowledge or hand labeled data.

In urban studies, Sakaki et al. (2012) explored transportation information detection from twitter. Balduini et al. (2013) provided a city scale event detection method which links tweets with RDF data streams to support continuous query for burst detection. Anantharam et al. (2015) proposed a supervised approach based on automated creation of a training set, with a focus on traffic related events. Zhou et al. (2016) adopted an unsupervised approach to extract city events from Twitter streams, contributed a location entity recognition model to obtain the precise location of related events and a qualitative estimation of the impact of the detected events.

Methods for event identification and extraction from social media are also broadly classified as generic (Ritter et al., 2012) or event-specific (Ritter et al., 2015). For city planners and authorities, very often it is needed to follow some specific types of events. To address such needs, we take an event-type specific approach and try to find if certain types of city events are visible on Instagram. As most of the reported studies have focused on Twitter data, our study exploring an Instagram dataset adds a new perspective.

Our approach is inspired by the more recent weakly supervised methods for detecting and extracting a particular kind of events from tweets (Ritter et al., 2015; Jain et al., 2016). We combine the use of primitive event specific knowledge (event name, type, location), with its expansion through semantic similarity analysis using word2vec to obtain additional related terms and phrases. We apply LDA topic modelling and visualization to the extracted event-relevant data, to help us discover hidden event topics and indicators. Our results help to shed light on the presence or absence of certain specific types of city events on Instagram.

2 DATA AND METHODS

Our dataset contains a 3-month collection of Instagram posts and comments from publicly accessible Instagram accounts in the Helsinki metropolitan region during the summer 2016[2]. Language detection found Posts and Comments in 47 languages, with the majority being Finnish (169,826) and English (111,157). Posts and comments are con-

[2] Data were collected by the Digital Geography Lab at University of Helsinki (https://www.helsinki.fi/en/researchgroups/digital-geography-lab)

sidered separate entries/documents, but can be merged easily at analysis. Hashtags are kept as effective content for analysis. The dataset is completely unlabeled. Table 1 gives an overview of the dataset.

Corpus	English	Finnish
Number of total entry	111,157	169,826
Number of posts	75,354	88,877
Number of comments	35,803	80,949
Number of words	1,340,185	1,359,174
Vocabulary size	140,275	260,902
Longest post in words	1,985	2,087
Average length	111	89

Table 1: Dataset overview

Event Types	Key word	Location/ Venue
Urban festival	festival	seurasaari
Sports	sport	suomenlinna
Safety	safety	
Security	security	

Table 2: Initial Seed Terms

Pre-processing cleans the data (removing noise symbols), runs POS tagging, removes stop-words (customized list), generate n-grams. N-grams with prepositions and conjunctions at beginning and end are discarded. Words in n-gram phrases are combined to form one word, to make the text more consistent with hashtags. For unigrams, only nouns and verbs are kept.

Our method for focused city events detection includes three major components: defining targeted event types and seed terms; expanding the seeds based on word2vec model; retrieving relevant Instagram posts and LDA topic modelling on the retrieved posts collection.

2.1 Targeted Events and Seeds

At the starting of the process, user's knowledge about a given city or region can help easily define the targeted events or event locations of interest, in the form of 1-2 keywords or keyphrases associated with the event type or location. These terms are strong event indicators from users' point of view[3].

Name of event types would be one of the best strong indicators. We are particularly interested in urban festivals, sports events, as well as public safety and security events. Table 2 lists the initial seed terms for the different types of city events. Note that the location and event type should be read as independent of each other.

2.2 Seed Expansion

Next, the initial seed terms/phrases are used to acquire additional event cues and indicators. This is done by using word embedding to find new terms that have high semantic similarity and relatedness with the seeds.

[3] an event indicator refers to "any word that is highly relevant to an event, can be used as a seed term to detect event mention in text" (Jain et al, 2016),

For this purpose, we tested with two word2vec models: (1) our own model trained using cleaned original sentences of the Instagram posts and comments as input, excluding all hashtags; (2) pre-trained word vectors from Facebook fasttext (https://github.com/facebookresearch/fastText). Our testing showed that the fasttext word vectors tend to expand the seeds to more generally related terms, whereas our own word2vec model helps to expand the seeds to related terms in a more specific context (such as things more specific to the city). We thus consider using our own word vectors in this study to acquire relatively strong indicators. Later on, fasttext word vectors could be used to obtain relatively weaker event indicators.

2.3 Retrieving Event Relevant Posts, LDA Topic Modelling

We incorporated two approaches for retrieving event relevant posts from the data: the single keyword based method, and the collective method that combine the initial seeds together with newly acquired expansions to search for and extract set of event relevant posts that are likely to contain mentions of the target event and other indicators.

We perform LDA topic modelling analysis (Blei et al., 2003; Hoffman et al., 2010; Blei, 2012) and visualization using LDAvis (Sievert and Shirley, 2014) on the event relevant posts. We manually identify some interesting types of events or their seeds (instances of a certain types of events), which can be used again as cues for searching relevant posts/comments. LDA modeling analysis in particular helps to bring up multiword hashtags that are not used during training of word vectors.

3 EXPERIMENTS

We experimented with the event types and initial seeds shown in Table 2.

3.1 Festival Events

The initial seed for festival events is the single word “festival”. Keyword based search for relevant posts collected 5238 posts. A 10-topic LDA model (Figure 1) reveals some interesting information:

- The term “flowfestival” is the most frequent word, followed by “flow” and “festival”.
- There are some other very interesting related words: music, exhibition, interior, fashion, florenceandthemachine, summersound, asuntomessut, ruisrock, weekendfestival, tuska. These would be highly valued expansions of event indicators - the first four represent event subcategories and others represent names of specific events.
- In addition, it’s interesting to notice temporal terms such as wknd, weekend, Sunday, Monday, night.
- The most prevalent festival events seem to be related to “music” and “weekend”.

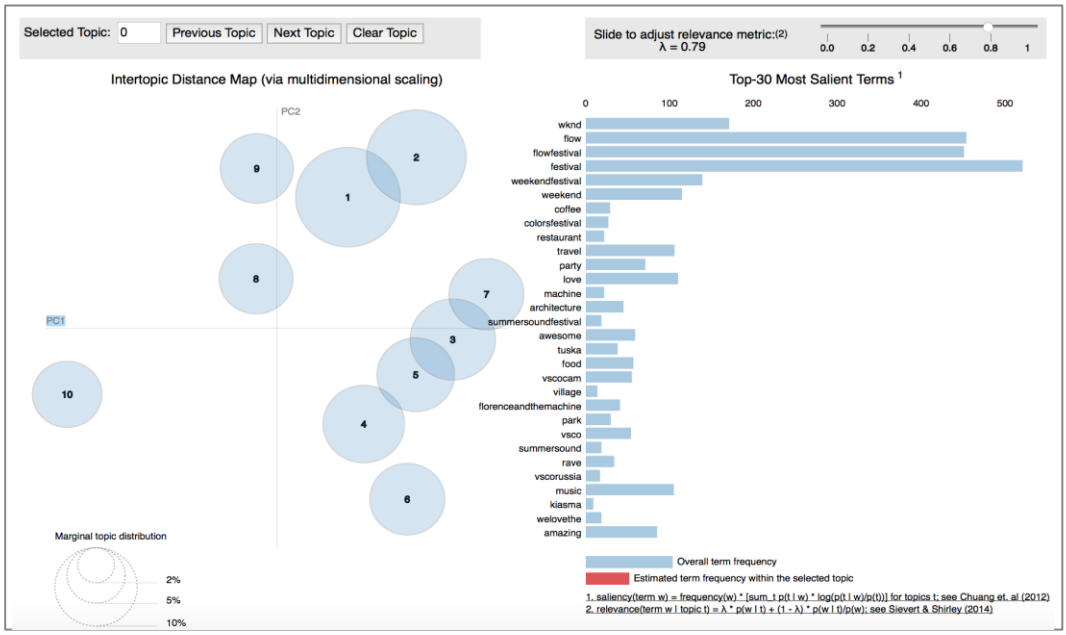

Figure 1. 10-topic lda model, seed "festival"

Using word2vec model however, the 5 most similar content words to "festival" are: 'tuska', 'throwback', 'crayfish', 'flow', 'linnanmäki'. These are related words for "festivals", with some overlapping with what we discovered from the 10-topic lda model. However, comparing with the top terms revealed from earlier lda topic model, these additions may tend to break context boundaries much easier.

When using the seed term "festival" together with expansion to the 5 related terms found using word embedding, a 10-topic lda model of the retrieved Instagram posts also reveals some interesting information:

- Some of the most frequent terms include "flowers, flow, flowfestival, tuska, blackworkers, blacktattooart, fashiontattatoo, helsinkiart, artfashion, linnanmäki". This seems diverted a bit far from the festival events.
- The most prevalent topics tend to capture more topic terms than event terms. Many festival events discovered in the last step of lda modeling become hidden.

3.2 Sport Events, Safety and Security Events

We started exploring sport events with the single seed term "sports", which is then expanded using word embedding based top similar words (again we consider them not very high quality). Keyword based search and expanded search obtained 1547 and 2776 Instagram posts respectively. Figure 2 shows the lda model based on keyword search.

In both cases the lda models revealed sports related topic terms such as "sports, watersports, powerboat, waterfun, running, training, sailing, workout, fitness" as well as

much travel related content, but not very many public sport events, except “helsinkicityrun” was visible in the keyword based approach.

Searching for Safety and Security related posts did not find out much real relevant content - maybe this is a good thing.

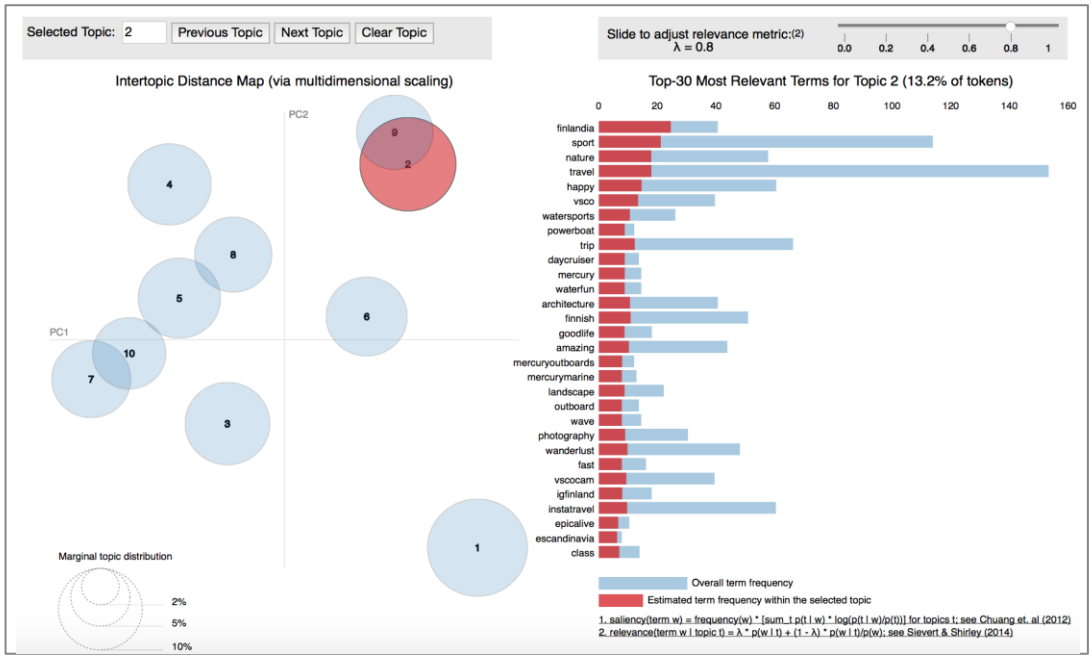

Figure 2. 10-topic lda model, seed ”sport”

3.3 Location/Venue Specific Events

Seurasaari and Suomenlinna are nature park and popular tourist attractions. Simple keyword based search only resulted in 93 posts that has mentions of Seurasaari. The LDA visualization was able to reveal some highly relevant topic terms such as “view, wood, sunset, pavillion, museum, island, landscape, nature, green, bird, naturelovers”, but rarely a mention of city events, except “midsummer”.

Keyword search with “Suomenlinna” collected much more posts (1077). Nonetheless, LDA visualization mostly found highly relevant topic terms such as “sveaborg, island, nature, travel, ferry, rocks, seafortress, seagulls, rocks, history, landscape, picnic”. City events related with the location is basically invisible. The interesting result from this part of the experiment is that the word2vec model was able to find highly similar terms “fortress', 'island', 'käpylä', 'senate', 'sea', 'baltic', 'sveaborg', 'uspenski', 'cathedral'. The LDA modeling results from expanded set of relevant posts provides us richer and much relevant information about Suomenlinna than the keyword based approach, as is shown in Figure 3.

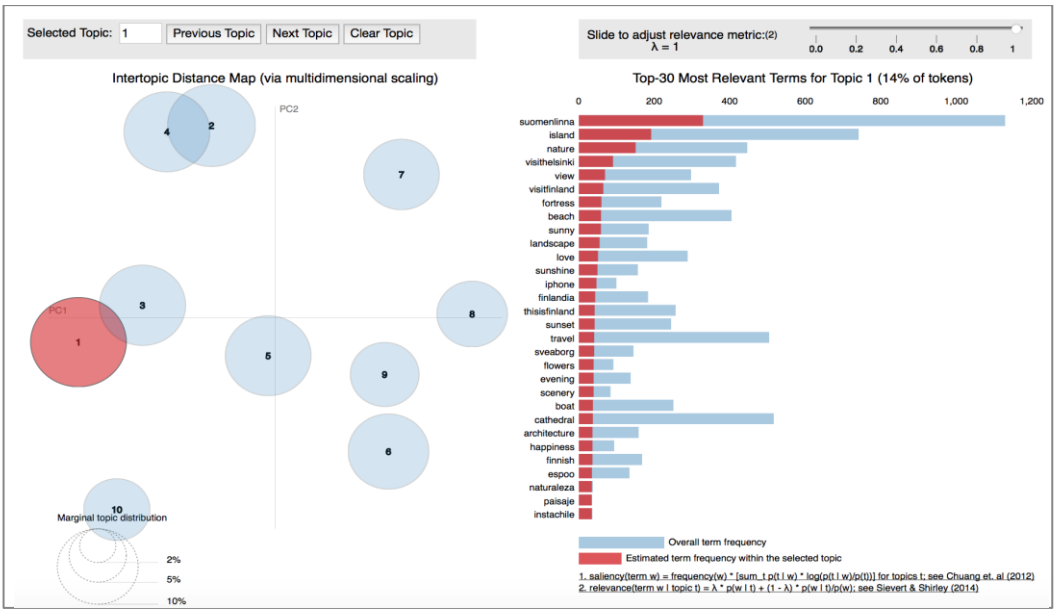

Figure 3. 10-topic lda model, seed "suomenlinna" with word2vec expansion using top ten most similar

4 CONCLUSIONS

In this study, we developed a system that aims to detect specific types of city events from Instagram posts. Our experiments help us to obtain some insights into content related with city events on the Instagram platform.

Overall, for the types of events we explored, it is possible to identify event mentions and event indicators with keyword based search for relevant posts followed by lda topic modelling and visualization. The expanded set of event indicators showed some positive effect only on the event type "Suomenlinna". In cases when the seed terms are more generic in nature such as "festival, sports, safety", the expansion can quickly bring in much irrelevant content or noise. In other cases, there is simply a rare presence of the concerned events in the Instagram data.

This study is our beginning step towards an enriched understanding of cities and regions via social media analysis. Automatic identification of a particular class of events in social media will enable downstream applications to help us better understand our city and region. Our system at this stage is not fully automatic, and best results come from combing user insights with system usage. Next step it will be interesting to compare with automatically learned event indicators similar in Jain et al. (2016), to integrate a named entity recognizer, especially to incorporate the Time indicators for events.

REFERENCES

Anantharam Pramod, Payam Barnagh, Krishnaprasad Thirunarayan and Amit Sheth. 2015. Extracting City Traffic Events from Social Streams. *ACM Transactions on Intelligent Systems Technology* 6(4): pp.1-27.

Atefeh F and W Khreich. 2015. A survey of techniques for event detection in twitter, *Computational Intelligence* 31 (1): pp. 132-164.

Balduini, Marco, Emanuele Della Valle, Daniele Dell'Aglio, Mikalai Tsytsarau, Themis Palpanas and Cristian Confalonieri. 2013. Social listening of City Scale Events using the Streaming Linked Data Framework", *Proceedings of the 12th International Semantic Web Conference - Part II*. 2013. Springer-Verlag New York, Inc.

Blei D, A. Ng and M. I. Jordan. 2003. Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems:* pp. 601-608.

Blei D. 2012. Probabilistic Topic Models. *Communications of the ACM* 55(4): pp. 77-84.

Dong Xiaowen, Dimitrios Mavroeidis, Francesco Calabrese, and Pascal Frossard. 2015. *Multiscale Event Detection in Social Media. Data Mining and Knowledge Discovery* 29, no. 5: pp. 1374–1405.

Hoffman M, D. Blei and F. Bach. 2010. Online learning for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems* 23: pp. 856-864.

Jain Ajit, Girish Kasiviswanathan, and Ruihong Huang. 2016. Towards Accurate Event Detection in Social Media: A Weakly Supervised Approach for Learning Implicit Event Indicators. *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 23–30, Osaka, Japan.

Li Jiwei, Alan Ritter, Claire Cardie and Eduard Hovy. 2014. Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Alan Ritter, Mausam, Oren Etzioni, Sam Clark. 2012. Open domain event extraction from twitter. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining:* pp. 1104-1112.

Ritter Alan, E Wright, W Casey and T Mitchell. 2015. Weakly supervised extraction of computer security events from twitter, *Proceedings of the 24th International Conference on World Wide Web:* pp. 896-905.

Sakaki Takeshi, Makoto Okazaki and Yutaka Matsuo. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *WWW2010*, Raleigh, North Carolina.

Sasaki Kenta, Shinichi Nagano, Koji Ueno and Kenta Cho. 2012. Feasibility Study on Detection of Transportation Information Exploiting Twitter as a Sensor. *AAAI Technical Report* WS-12-04, 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org).

Sievert C. and K. Shirley. 2014. LDAVis: A Method for Visualizing and Interpreting Topics. *ACL Workshop on Interactive Language Learning, Visualization and Interfaces*, Baltimore.

Tasse Dan and Jason I. Hong. 2014. Using Social Media Data to Understand Cities. *Proceedings of NSF Workshop on Big Data and Urban Informatics.*

Zhou Y, De, S. and K Moessner. 2016. Real world city event extraction from Twitter data streams. *International Workshop on Data Mining on IoT Systems (DaMIS16)*, London, UK.

Mikolov Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR.*

Mikolov Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS.*